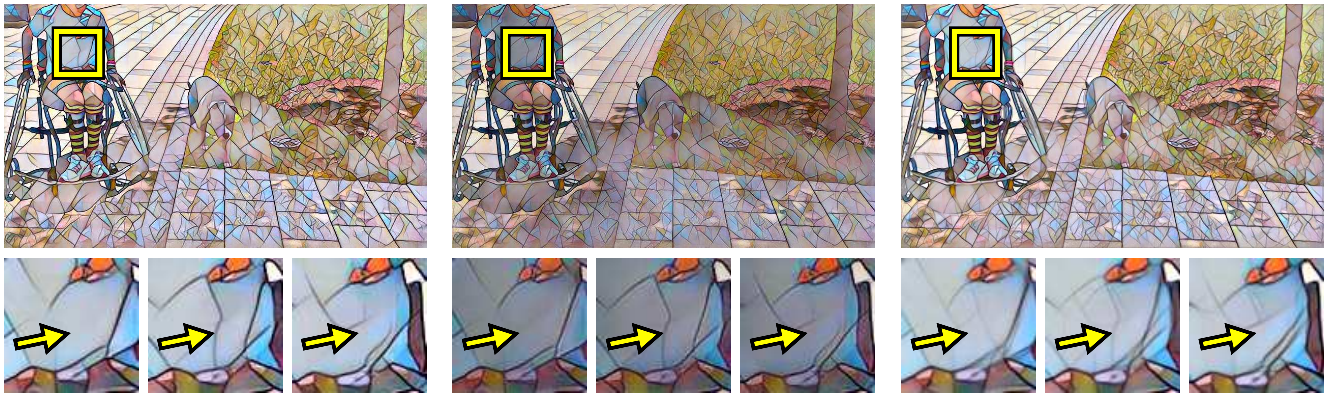
HyperCon: Image-To-Video Model Transfer for Video-To-Video Translation Tasks – Supplementary Materials

1. Additional Results for Style Transfer

In Figures 1 and 2, we compare HyperCon's predictions to the baselines for additional style transfer examples. We see that HyperCon produces more temporally consistent predictions than the baselines.

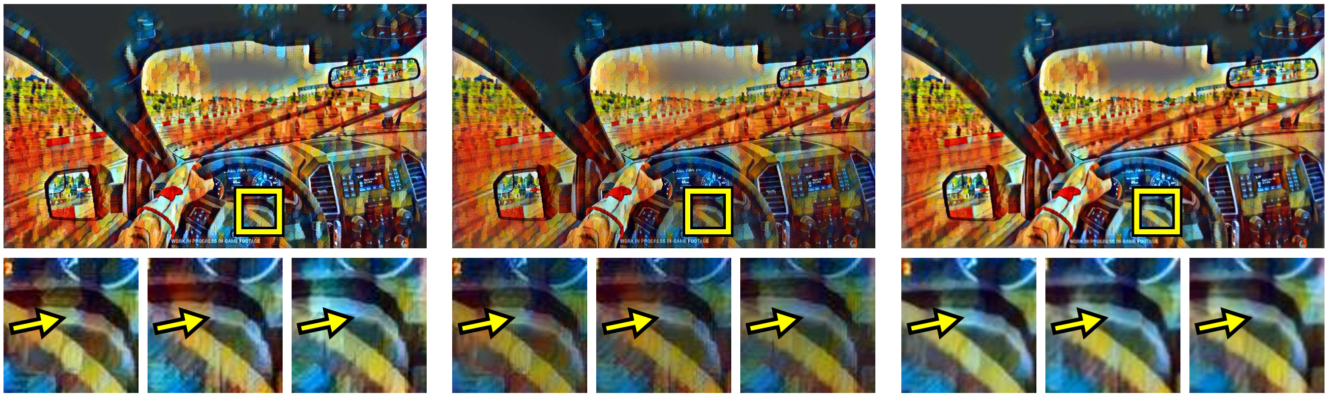


(a) FST [1]

(b) FST-vcons [2]

(c) HyperCon (ours)

Figure 1: Comparison of flickering artifacts on the *mosaic* style. The baselines (FST and FST-vcons) generate a vertical mark that appears for one frame, whereas HyperCon almost completely removes it.

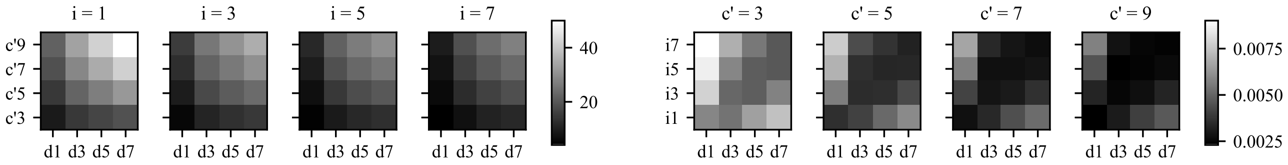


(a) FST [1]

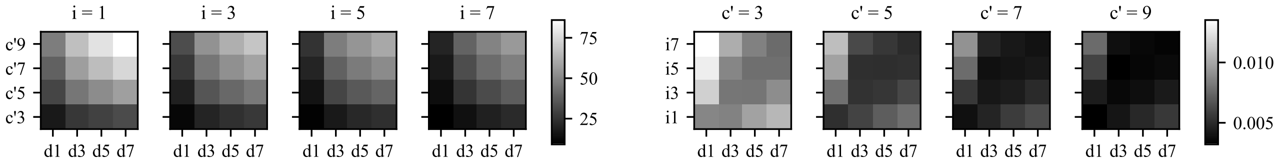
(b) FST-vcons [2]

(c) HyperCon (ours)

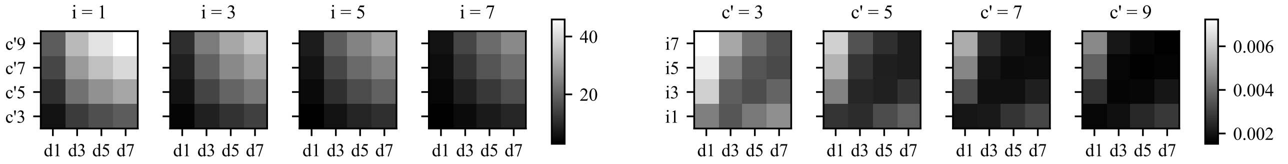
Figure 2: Comparison of flickering artifacts on the *rain-princess* style. The baselines (FST and FST-vcons) frequently changes the colors within the indicated area, whereas HyperCon predicts a stable color across frames.



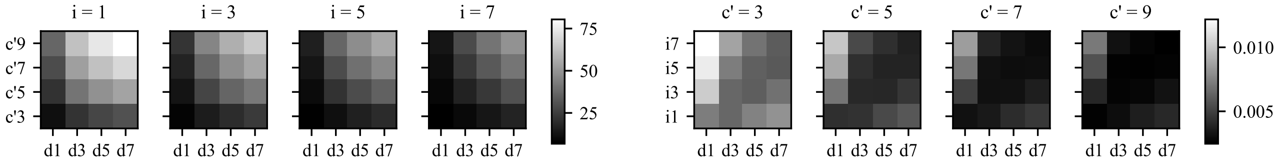
(a) *rain-princess*, DAVIS train/val



(b) *mosaic*, DAVIS train/val



(c) *rain-princess*, ActivityNet val



(d) *mosaic*, ActivityNet val

Figure 3: Ablative analysis plots. Left plots show FID (style adherence) and right plots show E_{warp} (temporal consistency). Lower is better.

1.1. Ablative Analysis

Figure 3 includes quantitative plots for our ablative analysis for all styles and validation sets; we observe similar trends across all styles and validation sets. Turning to qualitative results in Figure 4, we confirm that a low FID indicates strong style adherence and that a low E_{warp} indicates strong temporal consistency. However, strictly minimizing one or the other does not yield the most visually satisfying results. For instance, the hyperparameters that minimize FID exhibit the same flickering artifacts as frame-wise style transfer (FST)—observe that the region next to the cow's foot suddenly changes from blue to orange in the final frame for both FST and the lowest FID model. Meanwhile, the hyperparameters that minimize E_{warp} yield blurry predictions, which is the result of overly smearing several intermediate predictions across frames. Our final model produces consistent tones without overblurring, indicating that our selection strategy sensibly compromises between the two objectives.

1.2. Style Transfer Survey Design

To gauge the quality of our style transfer results, we have conducted a human evaluation on Amazon Mechanical Turk[1] (AMT). The motivation behind this study is primarily to ensure that the quantitative metrics that we use in the main paper correlate with human judgement at scale.

For our study, we developed custom web interfaces shown in Figure 5. The displayed videos are synchronized at all times, and playback can be toggled by pressing the "Play/Pause" button. Users can choose to display one or multiple videos at a time; if one video is selected, it is centered at the top of the page, and if multiple videos are selected, any non-selected video

---

[1] `https://www.mturk.com/`

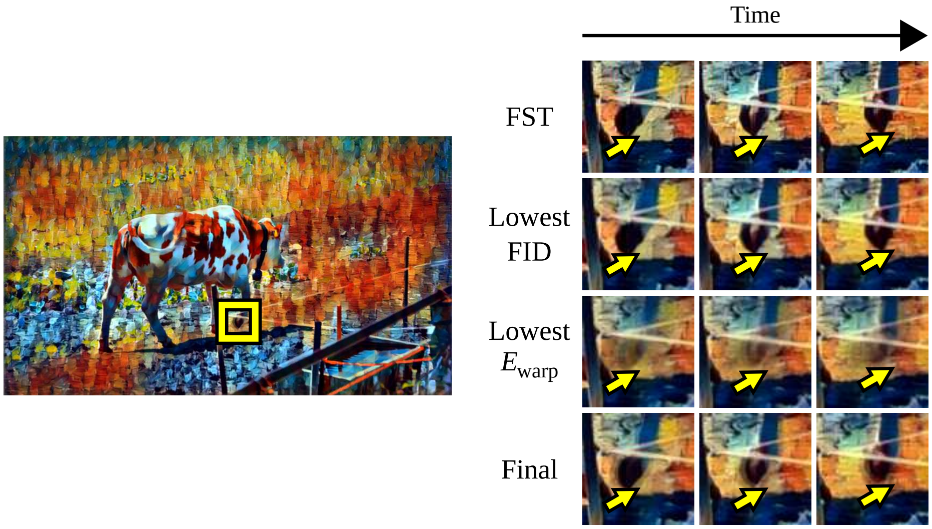


Figure 4: Qualitative comparison between frame-wise style transfer (FST) and ablative variants of HyperCon. The "Lowest FID" model reproduces flickering artifacts (*e.g.*, the changing tone near the arrow), while the "Lowest E_{warp}" model overly blurs predictions. Our final model adheres to the intended style without overly blurring predictions.

is replaced with a black frame. Playback does not reset when different videos are selected, so users can effectively perform A/B-style viewing without losing their place in the video.

Using these interfaces, we asked subjects two questions:

- "Which video is more appealing?" (preferred)
- "Which video looks more to 1?" (style adherence)

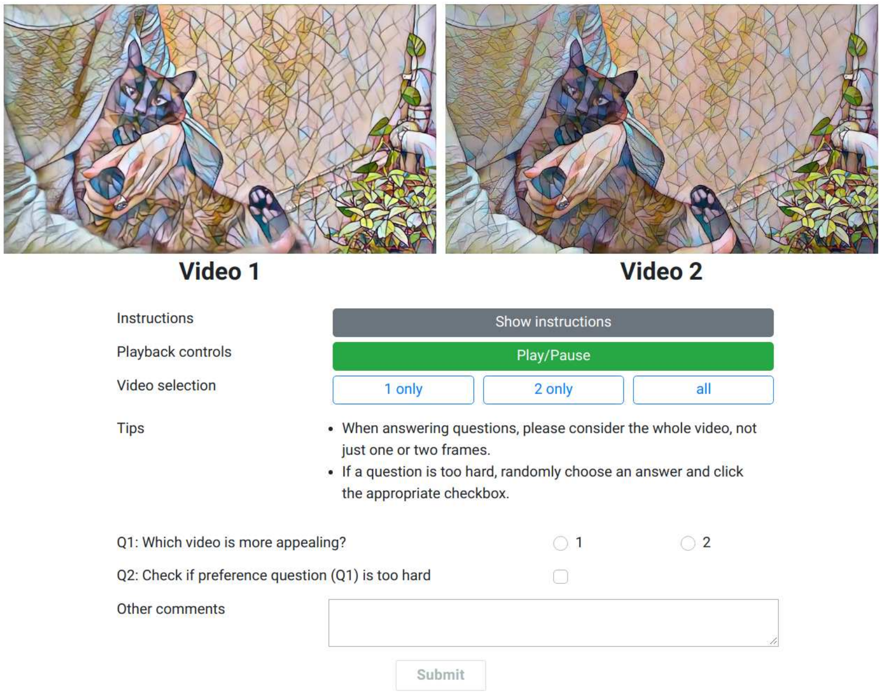
For the "preferred" question, we intentionally omitted instructions on what exactly makes a video more appealing since such qualities are inherently ambiguous. As for the "style adherence" question, we framed it in terms of comparing the pairs of videos under evaluation to the frame-wise stylized video as reference; we determined this to be the most concrete, unambiguous way to define the intended video stylization for subjects.

For each combination of style and source video, we randomly selected which method (between HyperCon and the FST-vcons baseline) would appear as the first and second video under evaluation (*i.e.*, videos 1 and 2 in Figure 5a and videos 2 and 3 in Figure 5b). This forces subjects to watch all videos carefully when answering each question. For the "style adherence" question, the frame-wise stylized video (FST) always appeared as video 1.

We also provided a checkbox to allow the subject to indicate that a question was difficult. We had considered an alternative survey formulation that only asked one question, but offered three responses (*e.g.*, "1", "2", or "neither" for the "preferred" survey), but opted for the two-question approach to ensure that rejecting the null hypothesis would lead to an interpretable result. We observed that the checkbox was selected for 5.08% of all responses to the "preferred" question and 6.61% of all responses to the "style adherence" question, indicating that the two methods were clearly distinctive in most cases.

We paid \$0.10 USD per task, which equates to \$12 USD/hour if each survey takes 30 seconds to answer. A total of about 150 subjects participated in our experiments. Surveys were generated as separate Human Intelligence Tasks (HITs), so it is not necessarily the case that any given subject saw both questions for any or all videos/styles.

In terms of filtering subjects, we did not utilize any demographic-based filters; instead, we created a qualification test utilizing an interface equivalent to Figure 5b. The videos used in this test were derived from a source video not included in our DAVIS and ActivityNet test videos, and had a simple variable darkening filter applied to each frame (*i.e.*, multiplying all RGB values by some constant). Examples of the effects we applied include darkening all frames by the same amount, darkening each frame by a random amount, and not darkening any frame. For each video triplet, we always included one of the possible answers as the reference; this effectively forced subjects to match the reference video with one of the other two. We also ensured that the odd video out was blatantly different to make the qualification test straightforward. We generated 10 triplets of qualification test videos, and required subjects to answer correctly for at least 8 of them before working on our tasks. For our final survey, we solicited 5 responses for each combination of source video, style, and question.

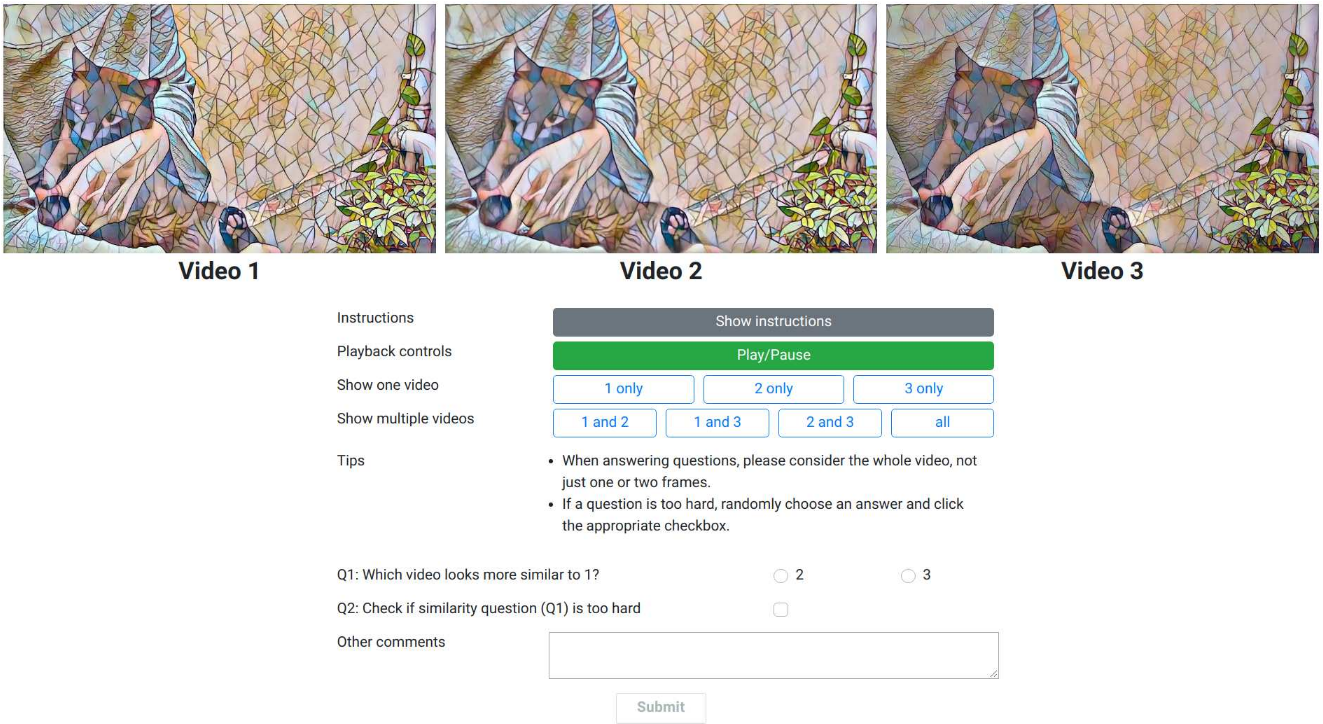


- When answering questions, please consider the whole video, not just one or two frames.
- If a question is too hard, randomly choose an answer and click the appropriate checkbox.

Q1: Which video is more appealing? ○ 1 ○ 2

Q2: Check if preference question (Q1) is too hard ☐

(a) Preferred



- When answering questions, please consider the whole video, not just one or two frames.
- If a question is too hard, randomly choose an answer and click the appropriate checkbox.

Q1: Which video looks more similar to 1? ○ 2 ○ 3

Q2: Check if similarity question (Q1) is too hard ☐

(b) Style adherence

Figure 5: Examples of the interfaces used for the “preferred” and “style adherence” surveys.

	mosaic		rain-princess	
	Preferred	Style adherence	Preferred	Style adherence
FST-vcons [2]	42.7%	20.9%	44.5%	23.2%
HyperCon (ours)	**57.3%**	**79.1%**	**55.5%**	**76.8%**
p-value	5.53×10^{-7}	0*	1.54×10^{-4}	0*

(a)

	DAVIS 2017 (60 source videos)				DAVIS 2019 (60 source videos)				ActivityNet (116 source videos)			
	mosaic		rain-princess		mosaic		rain-princess		mosaic		rain-princess	
Method	Preferred	Style adherence	Preferred	Style adherence	Preferred	Style adherence	Preferred	Style adherence	Preferred	Style adherence	Preferred	Style adherence
FST-vcons [2]	137	45	128	68	133	43	147	63	234	159	250	143
HyperCon (ours)	**163**	**255**	**172**	**232**	**167**	**257**	**153**	**237**	**346**	**421**	**330**	**437**
Total	300	300	300	300	300	300	300	300	580	580	580	580

(b)

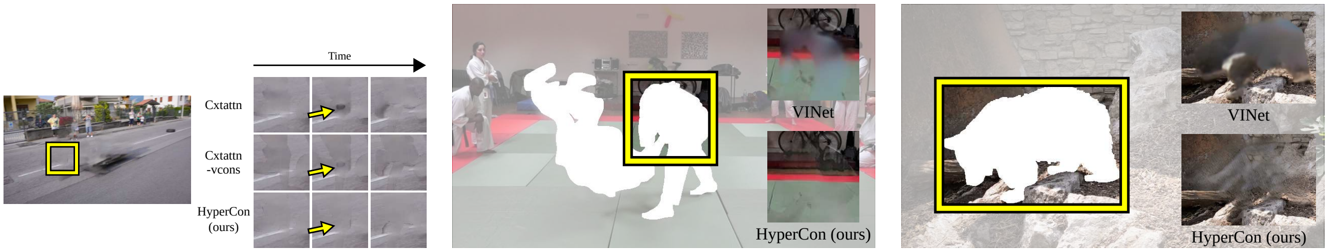
Table 1: Human evaluation of style transfer quality. (a) For each style, we list how often subjects favorably select each method across all videos of that style, as well as the p-value of the corresponding χ^2 test. 0* indicates a p-value less than 1×10^{-10}. In all cases, subjects select HyperCon (ours) significantly more often than FST-vcons [2]. (b) A detailed breakdown of responses for each style and dataset.

1.3. Detailed Style Transfer Survey Results

We provide an aggregate view of our human evaluation results in Table 1a. For the "preferred" question, subjects select HyperCon more often than FST-vcons, enough to reject the null hypothesis that subjects select each method with equal probability—in short, they prefer our predictions over those of the baseline. Furthermore, for the "style adherence" question, which asks for the video that is more similar to the reference style video, subjects also select our HyperCon method significantly more often; this means that our method is better at preserving the intended style than FST-vcons. This matches the conclusion made from our method's lower FID scores, and indicates that FID correlates with video quality and style adherence. In Table 1b, we provide a detailed breakdown of responses for each question and dataset.

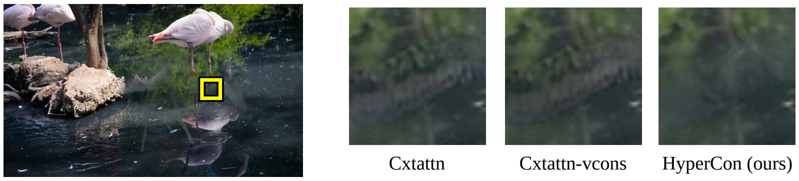
2. Additional Qualitative Results for Inpainting

This section provides additional qualitative results for the inpainting task on DAVIS 2017 training/validation videos. In Figure 6, we depict additional examples where HyperCon reduces flickering and boundary distortion effects, as well as produces more realistic texture, compared to the baselines. Next, in Figure 7, we show examples in which the image-to-video model transfer baselines produce checkerboard artifacts and HyperCon does not. Finally, in Figure 8, we compare our method to Cxtattn-vcons in cases where Cxtattn-vcons fails to make a prediction that blends well with the known pixels due to the hue shift problem.

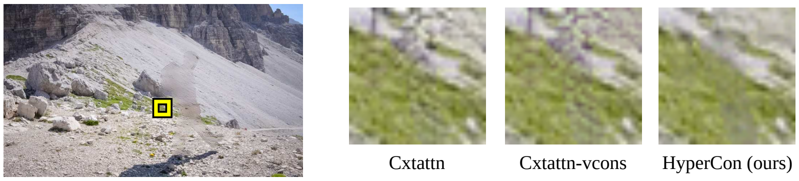


(a) Flickering (b) Boundary distortion (c) Texture comparison

Figure 6: Additional examples of flickering and boundary distortion effects from the baselines versus HyperCon for video inpainting. (a) HyperCon reduces artifacts better than the image-to-video model transfer baselines. The gray circle is apparent in the Cxtattn and Cxtattn-vcons prediction, but not in the HyperCon one. (b) VINet fails to connect the boundary of the mat in the background, whereas HyperCon successfully does connect the boundary. (c) VINet produces overly smooth textures that do not blend in well with the surrounding region, whereas HyperCon produces more realistic textures.

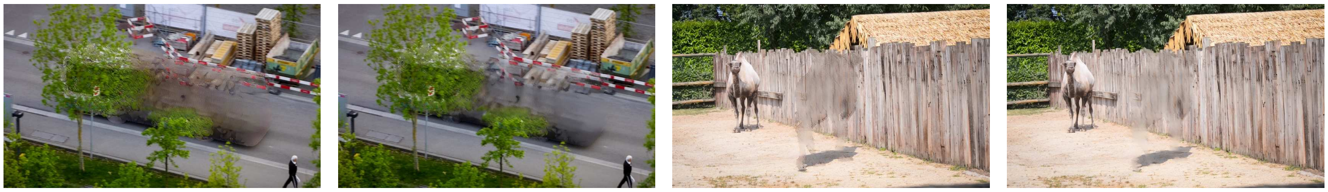


(a) Flamingo



(b) Hike

Figure 7: HyperCon generates fewer checkerboard artifacts than the baselines due to their instability across frames.



(a) Cxtattn-vcons (b) HyperCon (ours) (c) Cxtattn-vcons (d) HyperCon (ours)

Figure 8: Cxtattn-vcons distorts the hue of the inpainted region; HyperCon does not. As a result, our HyperCon predictions blend in more convincingly with the surrounding area.

References

[1] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *European Conference on Computer Vision*, 2016.

[2] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning Blind Video Temporal Consistency. In *European Conference on Computer Vision*, August 2018.